ATLANTIS PRESS

Advances in Economics, Business and Management Research, volume 72

12th International Conference on Business and Management Research (ICBMR 2018)

Zero-Inflated Poisson Regression Analysis on Frequency of Health Insurance Claim PT. XYZ

Rahmaniar Dwinta Kusuma
Faculty of Economics and Business
Universitas Indonesia
Depok, Indonesia
rahmaniar.dwinta@ui.ac.id

Yogo Purwono
Faculty of Economics and Business
Universitas Indonesia
Depok, Indonesia
yogo.purwono@ui.ac.id

***Abstract*—Modeling data count is an important thing in various fields. For this purpose, Poisson regression models are often used. However, in this model there is an assumption of equidispersion data where the mean value equals the value of the variance. In fact, this assumption is often violated in the observed data. In real data, the value of variance actually exceeds the mean (overly dispersed) value with the cause of the overdispersion depending on many situations. When the overdispersion source is exceeds zero (excess zero), then a more suitable model to use is the Zero-inflated Poisson regression model. In this paper, after the framework of Poisson regression and the Zero-inflated Poisson regression is reviewed then the model is adjusted to the claim frequency data in a private health insurance scheme where the frequency of claims is overly dispersed because of the number of zeros in the data set. Then Vuong's test is done to compare the two models and obtain the result that the Zero-inflated Poisson regression is more suitable for modeling the frequency data of PT.XYZ Health Insurance claims.**

***Index Terms*—overdispersion, zero-inflated Poisson regression, Vuong's test**

I. Introduction

Currently the development of the insurance industry in Indonesia is increasing rapidly. Insurance conditions continue to grow from year to year due to several influencing factors such as demography, economic expansion and new regulations that continue to drive towards acceleration. In addition, many Indonesians are now well aware of the importance of insurance. This is related to modern society who consider that insurance is a need for self and property protection from calamities that can occur in the future. The fundamental purpose of an insurance company is to calculate the appropriate insurance and premium prices for the insured to cover certain risks. The method used by actuaries to calculate premium is pure premium method, where this premium is the average premium that must be collected to pay only for losses [1]. The calculation uses historical data by dividing losses by the number of exposures (a period of time that states the validity of an insurance policy, where the status of the insurance policy is active). Pure premium is calculated by multiplying conditional expectations of claim frequency by the estimated claim costs (severity of claims). In this method the model for claim frequency and severity must be distinguished.

Claim frequency is a discrete, dependent variable in the form of data count. Count data is widely used in various fields, such as insurance, public health, epidemiology, psychology, and many other research areas. When the response variable is in the form of data count, the regression analysis commonly used is Poisson regression analysis. In general, Poisson distribution is a realistic model for various random phenomena as long as the value of the Poisson random variable is a non-negative integer. As mentioned by Boucher and Guillen, regression analysis count allows identification of risk factors and predicted expected frequency of claims with presence of risk characteristics [2]. Poisson regression is a special case of GLM (General Linear Model) which was first developed by Nelder and Wedderburn [3]. Although this model has many advantages, it is emphasized that in the Poisson distribution there are significant limitations that limit its use. Poisson distribution has a basic assumption that variance value and mean value have the same value (equidispersion). Overdispersion (variance value is higher than mean value) that occurs in the data can be caused by several factors depending on the situation, one of which is due to the excess zeros in the data.

The most commonly used alternative distribution to overcome the overdispersion in the data is the Negative Binomial distribution introduced by Cameron and Trivedi [4]. Another log-likelihood comparison for Binomial Negative and Poisson distribution shows that additional parameter of Binomial Negative distribution can improve data match better than Poisson distribution. In fact, in some cases, data counts often have an excessive amount of zero data than should be expected in Poisson regression or Binomial Negative regression. So often zeros in the sample cannot be accommodated properly. So to solve the case of data with zero inflation like this requires the use of other more appropriate models. In this case the author discusses the application of Zero-Inflated Poisson Regression on the real data, that is, the frequency of claims data on health insurance.

II. Literature Review

Claim data is a discrete count data which is usually modeled using Poisson distribution. The application of Poisson distribution has a long history of General Insurance as a model for the number of claims. In addition, the Poisson distribution was introduced as a mixed distribution to describe non-homogeneous populations. Mixed Poisson distribution can

TABLE I: Description of health insurance data variables:

Variable	Categorical/Numeric	Description
Claim Frequency	Numeric	The number of claims that the insured has in each year
Gender	Categorical	1 for male and 0 for female
Age_30	Categorical	1 for the insured with an age less than and equal to 30 years, 0 for others
Age_40	Categorical	1 for the insured with a variation of age between 31-40 years, 0 for others
Age_60	Categorical	1 for the insured with a variation of age between 41-60 years, 0 for others
Marital Status	Categorical	1 for the insured with married status, 0 for the insured wit single status
Family Member	Numeric	Indicates the number of family members of the insured
Exposure	Numeric	Indicates the period of the insured at one year, the value varies from 0 to 1

be used to model a population consisting of a limited number of homogeneous sub-populations, in which the frequency of claims submitted by each policyholder can be modeled with a Poisson distribution.

As it is known that the assumption of equidispersion is usually not in accordance with the data in the insurance industry. The assumption that the population must be homogeneous when the Poisson distribution model being used cannot be fulfilled in the insurance data. The existence of heterogeneity problems arises because there are differences in the behavior of policyholders that cannot be observed by insurance companies. One consequence of heterogeneity is the emergence of overdispersion in data. This model cannot adequately model the claim frequencies because some assumptions cannot be fulfilled, as previously stated where the number of zeros in the empirical distribution far exceeds the number of zeros that are under the assumption of negative Poisson or binomial distributions. As a result, some researchers tried to modify the negative Poisson or binomial distribution to combine excess zeros in this distribution. The model that can handle the presence of excess zeros in the data observed is the Zero-inflated Poisson (ZIP) model [5]. Many research studies have been conducted before in assessing the application of the Zero-inflated Poisson structure model. This model was developed by Lambert [5] to handle zero-increase, then calculate data by combining two sources of zero values, namely “real zero” and “zero excess”. Various kinds of mixed distributions for Poisson distribution have been proposed as mentioned in his research entitled Zero-Inflated Regression, with an application to defects in manufacturing [5]. The study stated that the Zero-inflated Poisson model is a combination of Poisson distribution and logit distribution.

Basically the general form of zero-inflated Poisson regression model allows for a set of covariates for each parameter in the two models. Important risk factors related to claims behavior of policyholders are identified by the significance of the regression parameters. Greene conducted a study of zero-inflated models as modification of models from negative Poisson and binomial models [6]. There are many zero values on claim frequency data, usually the model leads to zero-inflated models. Therefore, it is necessary to test the model before the zero-inflated model is applied in the data. The researchers previously applied the chi-square test and the likelihood ratio test to test the goodness-of-fit of the model. In recent years the Zero-inflated score test provided by Van den Broek to test whether the sum of zeros is too large for a Poisson distribution to model data properly [7]. One of the advantages of this test is that it is not necessary to test the Zero-inflated Poisson model if the number of zeros owned is not excessive. Jansakul and Hinde add to the shortcomings of the Van den Broek study in the presence of tests for more general situations where the proportion of zeros in the data is left to depend on the covariate/with this test score, it can be determined whether the Poisson model is incorrect when compared to the Zero-inflated model [8].

Some recent research related to Zero-inflated Poisson is research conducted by Mouatassim and Ezzahid [9]. His research describes the comparison between the Poisson model and the Zero-inflated model and applies it to health insurance data sets. They concluded that Zero-inflated Poisson is more suitable than Poisson distribution for modeling operational risk frequencies. The most recent research applied to the health insurance industry is the application of Zero-inflated Poisson regression which is then compared with Poisson regression [9].

III. Research Methodology

A. Sample and Data

The data used in the study is secondary data obtained from health insurance company PT. XYZ. Frequency data/number of claims obtained from PT. XYZ is data from health insurance for 5 years (2011-2015). Claim frequency data is the dependent variable and the insured data which include gender, age, marital status, family members and exposure are independent variables. Then the data is processed into a claim frequency table, which consists of claim frequency and the insured data variables. The steps that can be taken are (1) based on each insured data calculated the frequency of claims made in each period; (2) classifying each variable of the insured data into a value of 0 or 1 according to the description attached to table 1; (3) data is processed using Easyfit and STATA software.

B. Zero-Inflated Poisson Regression

When the Poisson regression model is applied to real data, the results from the given model usually do not match the actual data based on its deviation value or Pearson chi-square. In the real data, the assumption of a standard Poisson model called equidispersion cannot be fulfilled, the data does not have the same mean and variance values and the variance values tend to be greater than the mean (overdispersion). If

the assumption of overdispersion is ignored and the standard Poisson regression is still applied, it causes an incorrect estimate of the standard error and an invalid significance level. One of the causes of overdispersion is excessive zero observations on the dependent variable.

One of the proposed or alternative methods for handling overdispersion is the Zero-Inflated Poisson (ZIP) regression model [8], where the distribution of ZIP regression model is a modification of Poisson distribution and logit distribution. With the possible value of Y being a nonnegative integer: 0,1,2,3, etc. This model was first proposed by Lambert with the application of defects that occur in the factory process [5]. In Zero-inflated Poisson regression, the dependent variable is mutually independent. According to Lambert [5], the Zero-Inflated Poisson model assumes a population or observation of two latent groups (unobserved). The whole model is a mixture of the probabilities of both groups that allow for overdispersion and zero excess that cannot be predicted by the standard Poisson model. An individual (observation unit) will enter in group A whose value is always zero (zero state) with probability p or will enter the group (non-zero state), where the value of zero and positive value is generated by a Poisson distribution function, with chances 1-p. So the probability functions for the zero and positive values that can be written in the equation is as follows:

$$Pr\left(Y=y_i\right)=\begin{cases}\theta_i\left(z_i\right)+\left(1-\theta_i\left(z_i\right)\right)Pois\left(\lambda_i;0\left|x_i\right.\right), \\ \text{jika } y_i=0 \\ \left(1-\theta_i\left(z_i\right)\right)Pois\left(\lambda_i;0\left|x_i\right.\right), \\ \text{jika } y_i>0\end{cases}$$

with:

- z_i is a covariate vector that defines the probability of θ_i
- $Poiss\left(\lambda_i;0\left|x_i\right.\right)=\exp\left(-\lambda_i\right)$
- $Poiss\left(\lambda_i;y_i\left|x_i\right.\right)=\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$, where λ is the mean and variance of the distribution.

Based on Lambert [5], θ_i can be modeled with Logit model (with γ is a vector of parameters):

$$\theta_i\left(z_i\right)=\frac{\exp\left(z_i'\gamma\right)}{\left[1+\exp\left(z_i'\gamma\right)\right]}$$

so the relationship model for λ and θ in the Zero Inflated Poisson regression model are:

$$\begin{aligned}\ln(\lambda) &= X\beta \\ logit(\theta) &= \log\frac{\theta}{1-\theta}=X\gamma\end{aligned}$$

It is assumed that $y_1, y_2, \cdots, y_n$ independent and θ_i is not related to λ_i. Then the likelihood function can be defined by:

$$\begin{aligned}L &= \prod_{y_i=0}\left[\theta_i\left(z_i\right)+\left(1-\theta_i\left(z_i\right)\right)\exp\left(-\lambda_i\right)\right] \\ &\quad \prod_{y_i\neq 0}\left[\left(1-\theta_i\left(z_i\right)\right)\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}\right]\end{aligned}$$

Furthermore, the log-likelihood function will be maximized by using EM algorithm (expectation maximization), this algorithm is one method of optimization. In the EM algorithm, the log-likelihood function that has missing data will be maximized. This algorithm estimates the expectations of 0 observation data in iteration and use expectations to estimate the parameters, where the iteration is carried out until it converges. Data that has a value of 0 in this problem is defined as an indicator variable $\delta=(\delta_1,\delta_2,\cdots,\delta_n)'$ where δ_i if y_i comes from zero state and $\delta_i=0$ if y_i comes from non-zero state. According to Lambert [5] and Hall DB and Shen J [10] the EM process can be written through 3 steps, E step, M stage for β and M for γ.

IV. Results

Below are tables of descriptive statistical values from the data, the value of descriptive statistics in the table is a value that has previously been changed to numbers 0 and 1 as mentioned in table I:

TABLE II: Descriptive statistics for numerical variables of PT.XYZ

Variable	Modus	Minimum	Maximum	Total
Claim Frequency	0	0	8	3633
Family Member	1	1	4	5802

Variable	Mean	Median	Minimum	Maximum	Total
Exposure	0,953	1	0	1	2673,9

TABLE III: Descriptive statistics for categorical variables of PT. XYZ

Variable	Category	Description	Percentage
Gender	0	Female	50,7%
	1	Male	49,3%
Age_30	0	Age $\leq$ 30	73,9%
	1	Others	26,1%
Age_40	0	Others	59,2%
	1	Age = 31–40	40,8%
Age_60	0	Others	63,6%
	1	Age = 41–60	36,4%
Marital Status	0	Single	49,4%
	1	Married	50,6%

Analysis using Zero-inflated Poisson regression is almost the same as standard Poisson regression. The analysis carried out in this study uses STATA software. Below is the parameter estimation result of the Zero-inflated Poisson regression model based on the results of the calculation:

Based on the estimated output of the Zero-inflated Poisson regression model above:

TABLE IV: Zero-inflated Poisson regression model estimation

		P(Y > 0)	*Inflate* P(Y = 0)
Intercept	Estimation	0,3143626	1,412,889
	Standard Error	0,2089019	0,5659153
	p-value	0,132	0,013
Gender	Estimation	-0,0189417	-0,1305908
	Standard Error	0,0367133	0,1080164
	p-value	0,606	0,227
Age_60	Estimation	0,3376694	-1,879,547
	Standard Error	0,696674	0,2824517
	p-value	0,000	0,000
Age_30	Estimation	-0,3270105	0,7609385
	Standard Error	0,1070958	0,3104704
	p-value	0,002	0,014
Age_40	Estimation	0,1349571	-0,2466598
	Standard Error	0,065	0,286852
	p-value	0,038	0,390
Marital Status	Estimation	0,5115299	-0,4220283
	Standard Error	0,0818647	0,207508
	p-value	0,000	0,042
Family Member	Estimation	-0,1644118	-0,3702375
	Standard Error	0,261148	0,0861873
	p-value	0,000	0,000
Exposure	Estimation	0,6954689	0,3946268
	Standard Error	0,1999657	0,4957177
	p-value	0,001	0,426

$$\begin{aligned} \ln(\lambda) &= -0,3270105X_2 + 0,1349571X_3 + \\ &\quad 0,3376694X_4 + 0,5115299X_5 - \\ &\quad 0,1644118X_6 + 0,6954689X_7 \\ logit(\theta) &= 1,412889 + 0,7609385X_2 - \\ &\quad 1,879547X_3 - 0,4220283X_5 - \\ &\quad 0,3702375X_6 \end{aligned}$$

The Zero-inflated Poisson regression model shows that variable Y, the claim frequency is influenced by six variables X, age_60, age_30, age_40, family member, marriage status and exposure with a significance level is α = 5%. Whereas for logit model there are 4 variables that have a significant influence, age_60, age_30, family member_, and marriage status_ with a significance level is α = 5%. The interpretation of the Zero-inflated Poisson regression model is almost the same as the Poisson regression model, which is by using the odds ratios of each coefficient. The parameter γ can be interpreted through the odds which are the ratio of the probability of something correct divided by an incorrect probability.

V. Discussion

One method that can be used in comparative test of the model between Zero-inflated Poisson regression and standard Poisson regression is the Vuong's test [11]. However, there are 3 other testing criteria that can be used as consideration in the selection of models, the likelihood ratio test and the AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) criteria.

Vuong test of zip vs. standard Poisson: z = 4.81 Pr $> z$ = 0.0000

Based on the tests performed, the values z = 4.81 and Pr $> z$ = 0,000 were obtained. The null hypothesis for testing Vuong's is a Zero-inflated Poisson regression according to the data used. The conclusion that can be obtained is that a significant z test indicates that the Zero-inflated Poisson model is more suitable than the ordinary Poisson regression model. Furthermore, the prediction of claim frequency for the next period will be calculated using the Zero-Inflated Poisson regression model obtained. Based on the calculation to determine the predicted value of the claim frequency used between the two models, the probability value for each model is calculated. Suppose that one of the insured policy data is known, based on the two models obtained the prediction of claim frequency using the Zero-inflated Poisson regression model is:

- Model $\ln(\lambda)$:
  Claim Frequency (λ) = - 0,3270105 (1) + 0,1349571 (0) + 0,3376694 (0) + 0,5115299 (0) - 0,1644118 (1) + 0,6954689 (1) = 0,204047. The claim frequency prediction result is 0.204047 because the number is still below 0.5, it will be rounded down to 0.
- Logit model (θ):
  Claim Frequency (θ) = 1,412889 + 0,7609385 (1) - 1,879547 (0) - 0,4220283 (0) - 0,3702375 (1) = 1,80359. The claim frequency prediction result is 1.80359 because the number behind the comma exceeds 0.5, it will be rounded up to 2.

After obtaining the probability value, the model used as the claim frequency prediction value is a model that has the highest probability value. The variable p is formed for each model with [4]:

- Predicted value of claim frequency based on $\ln$ model

(λ) = a1

- Predicted value of claim frequency based on logit model (θ) = a2
- pzero (probability value of the model $\ln(\lambda)$), where the value is:

$$pzero = \frac{\exp a2}{(1+\exp a2)} = \frac{\exp 1,80359}{1+\exp 1,80359} = 0,856$$

- pcount (probability value of logit model (model)), where the value is:

$$\begin{aligned} pcount &= \exp a1 * (1 - pzero) \\ &\exp 0,204047 * (1 - 0,856) \\ &0,176595 \end{aligned}$$

The results obtained for pzero are 0,856 and pcount is 0.1765. Then the highest probability value between the two models is the probability value for the model $\ln(\lambda)$ (pzero). The prediction value used is the value based on the $\ln(\lambda)$ model.

VI. Conclusion

There are several alternative methods that can be used to handle overdispersion in the data. In this study Zero-inflated Poisson regression model is used because of overdispersion in the data. Based on the results of data processing, obtained two models where in both models not all variables significantly influence the frequency of claims. The method used to compare Zero-inflated Poisson regression with Poisson regression is the Vuong's test. The results obtained show that the Zero-inflated Poisson regression model is more suitable for use in claim frequency data compared to Poisson regression. Another method that can be used to compare the two models is to use a comparison of the AIC, BIC and likelihood ratio tests. Later this model can be used as a material consideration in insurance company for determining the premium price, because the frequency of claims model is an important factor to be considered.

References

[1] D. Anderson, S. Feldblum, C. Modlin, D. Schrimacher, E. Schirmacher and N. Thandi, Practitioner's guide to generalized linear models: a foundation for theory, interpretation and application, Watson Wyatt, 2004.

[2] J. P. Boucher, M. Denuit, and M. Guillen, "Models of insurance claim counts with time dependence based on generalization of Poisson and Negative Binomial Distributions," Advancing the Science of Risk Variance, vol. 2, no. 1, pp. 135–162, 2008.

[3] J. A. Nelder and R. W. M. Wedderburn, "Generalised linear models," Journal of the Royal Statistical Society, Series A, vol. 135, pp. 370–384, 1972.

[4] A. C. Cameron and P. K. Trivedi, Regression Analysis of Count Data, New York: Cambridge University Press, 1998.

[5] D. Lambert, "Zero-inflated Poisson regression, with an application to defects in manufacturing," Technometrics Journal, vol. 34, pp. 1–14, 1992.

[6] W. Greene, "Accounting for excess zeros and sample selection in Poisson and negative binomial regression models," Working Paper EC-94-10, Department of Economics, New York University, 1994.

[7] J. V. D. Broek, "A score test for zero inflation in a Poisson distribution," Biometrics Journal, vol. 51, pp. 738–743.

[8] N. Jansakula and J. P. Hinde, "Score Tests for Zero-Inflated Poisson Models," Computational Statistics & Data Analysis, vol. 40, no. 1, pp. 75–96, 2002.

[9] Y. Mouatassim and E. H. Ezzahid, "Poisson regression and zero-inflated Poisson regression: Application to private health insurance data," European Actuary Journal, vol. 2, pp. 187–204, 2012.

[10] D. B. Hall and J. Shen, "Robust estimation for zero-inflated Poisson Regression," Scand Journal Statistics, vol. 37, pp. 237–252, 2010.

[11] Q. Vuong, "Likelihood ratio tests for model selection and non-nested hypotheses," Econometrica, vol. 57, pp. 307–334, 1989.